PRIFYSGOL ABERYSTWYTH UNIVERSITY

Aberystwyth University

Duplicate retention in signalling proteins and constraints from network dynamics

Soyer, O S; Creevey, C J

Published in:
BMC Evolutionary Biology

DOI:
10.1111/j.1420-9101.2010.02101.x

Publication date:
2010

Citation for published version (APA):
Soyer, O. S., & Creevey, C. J. (2010). Duplicate retention in signalling proteins and constraints from network dynamics. *BMC Evolutionary Biology*, *23*(11), 2410-2421. https://doi.org/10.1111/j.1420-9101.2010.02101.x

General rights
Copyright and moral rights for the publications made accessible in the Aberystwyth Research Portal (the Institutional Repository) are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the Aberystwyth Research Portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the Aberystwyth Research Portal

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

tel: +44 1970 62 2400
email: is@aber.ac.uk

Download date: 05. May. 2023

JOURNAL OF Evolutionary Biology

doi:10.1111/j.1420-9101.2010.02101.x

Duplicate retention in signalling proteins and constraints from network dynamics

O. S. SOYER* & C. J. CREEVEY†

*Systems Biology Program, College of Engineering, Computing and Mathematics and Physical Sciences, University of Exeter, Exeter, UK
†Animal & Grassland Research and Innovation Centre, Teagasc, Grange, Dunsany, Co. Meath, Ireland

Keywords:

comparative genomics;
duplication;
evolution;
fitness effects;
modelling;
protein family size;
response dynamics;
signalling network;
systems biology;
whole-genome duplication.

Abstract

Duplications are a major driving force behind evolution. Most duplicates are believed to fix through genetic drift, but it is not clear whether this process affects all duplications equally or whether there are certain gene families that are expected to show neutral expansions under certain circumstances. Here, we analyse the neutrality of duplications in different functional classes of signalling proteins based on their effects on response dynamics. We find that duplications involving intermediary proteins in a signalling network are neutral more often than those involving receptors. Although the fraction of neutral duplications in all functional classes increase with decreasing population size and selective pressure on dynamics, this effect is most pronounced for receptors, indicating a possible expansion of receptors in species with small population size. In line with such an expectation, we found a statistically significant increase in the number of receptors as a fraction of genome size in eukaryotes compared with prokaryotes. Although not confirmative, these results indicate that neutral processes can be a significant factor in shaping signalling networks and affect proteins from different functional classes differently.

Introduction

Both single-gene and whole-genome duplications (WGD) are well documented in various organisms (Brenner *et al.*, 1995; Zhang, 2003; Vogel & Chothia, 2006), and it is estimated that single-gene duplications happen at a rate similar to point mutations (Lynch & Conery, 2000; Lynch *et al.*, 2008). However, such high occurrence rates alone cannot explain the maintenance of duplicates over long time. For a duplicate to be maintained, it faces two evolutionary hurdles. First, it needs to increase in frequency in the population after its birth in one or few individuals. Second, there needs to be enough selective advantage for the duplicate so that both copies of the duplicated gene are maintained in face of deleterious mutations. One way to achieve such selective advantage would be for the duplicate to diversify from its origin. This is believed to occur through the accumulation of mutations leading to neofunctionalization (Walsh, 1995) and subfunctionalization (Force *et al.*, 1999; Lynch & Conery, 2000). There is substantial evidence for both processes (Evangelisti & Wagner, 2004; He & Zhang, 2005; Hughes & Liberles, 2007), with change in gene expression providing a major mechanism for duplicate retention (Huminiecki & Wolfe, 2004; Gu *et al.*, 2005; Duarte *et al.*, 2006; Tirosh & Barkai, 2007). On the other hand, initial fixation of a duplicate is less well understood. Redundancy (Nowak *et al.*, 1997; Wagner, 2000; Salathé & Soyer, 2008) and increased dosage (Cook *et al.*, 1998; Papp *et al.*, 2003) can lead to fixation through positive selection as shown in certain cases (Moore & Purugganan, 2003; Landry *et al.*, 2007). However, as indicated by these studies, such immediate selective advantage for a duplicate is only expected in certain gene classes. All other cases of duplicate fixation would occur through genetic drift.

For a duplicate to fix through genetic drift, its fitness effect has to be zero or below a critical threshold related to population size (Gillespie, 2004). Besides energetic costs (Wagner, 2005), the actual fitness effects of gene duplication (or loss after a WGD) will closely link to the

Correspondence: Orkun S. Soyer, College of Engineering, Computing and Mathematics and Physical Sciences, Harrison Building, North Park Road, University of Exeter, Exeter EX4 4QF, UK. Tel.: +44 (0)1392 723615; fax: +44 (0)1392 217965; e-mail: O.S.Soyer@exeter.ac.uk

© 2010 THE AUTHORS. *J. EVOL. BIOL.*
JOURNAL COMPILATION © 2010 EUROPEAN SOCIETY FOR EVOLUTIONARY BIOLOGY

function and structure of the protein product of the duplicated gene and its role in the larger biological system. For example, duplication of genes, whose products function as part of a complex, might have deleterious effects (Papp *et al.*, 2003; Deutschbauer *et al.*, 2005; Sopko *et al.*, 2006). For proteins involved in regulatory networks, theory suggests that most duplications would disrupt network dynamics and consequently the mediated gene expression patterns (Wagner, 1994). Similarly, for proteins involved in signalling networks, disruptions in network dynamics would be the main fitness effect associated with duplication. This is most readily imagined in single-celled organisms. For example, proper chemotaxis response in *Escherichia coli* requires the effector protein of the chemotaxis network to be in a certain concentration range (Cluzel *et al.*, 2000). Duplication of the effector (or any other protein in the network) could shift the network response out of this range and lead to loss of proper chemotaxis (Kollmann *et al.*, 2005).

Here, we investigate whether such dynamical effects of a duplication (or loss after a WGD) and consequently its fixation relate to its functional role in a signalling network. In particular, we consider four broad functional categories of signalling proteins as receptors, activators, deactivators and effectors. The latter three categories cover all intermediary proteins that relay the signal received at the receptor to an output protein such as a transcription factor or membrane channel, effectively translating the signal into a physiological response. To quantify the effects of duplicating a gene from these functional categories, one needs to systematically analyse the effect of duplication on response dynamics. However, there are not enough well-characterized signalling networks with experimentally verified reaction rates to achieve such a systematic analysis. To overcome this limitation, we rely here on a generic model of signalling networks that captures the response dynamics of such networks. Using this model, we create random networks and analyse the effects of duplication on response dynamics. By coupling such effects on response dynamics to organism fitness, we analyse how many duplications in a given functional class result in fitness effects below a critical fitness threshold (i.e. are neutral), as the level of selective pressure on the dynamics of the signalling network varies. To further support this theoretical analysis and overcome potential bias resulting from random models (Artzy-Randrup *et al.*, 2004), we also consider networks that are evolved *in silico* under selection for maintaining a given response dynamics. Analyses from both random and evolved networks give similar results and provide a general view of how neutral fixation of duplicates can be affected based on their functional role at network level. To see whether one of the main expectations of the model has any empirical support, we analyse the family size of signalling proteins in over 371 annotated genomes covering both eukaryotes and prokaryotes.

Methods

In the following paragraphs, we give a detailed description of the different models and approaches we used for the theoretical analysis and the empirical study.

Generic model of a signalling network

To capture the dynamics of signalling networks, we use a generic model similar to those developed previously (Heinrich *et al.*, 2002; Soyer *et al.*, 2006). In brief, we describe a network as a set of n interacting proteins. Each of these proteins is assumed to belong into one of four functional classes: receptor, activator, deactivator and effector. Proteins are assumed to have two states, an active (P_i^*) and an inactive state (P_i). Biologically, a protein can shift between such two states through phosphorylation, methylation or any other type of chemical or structural interaction mediated by another protein. To model such interaction, we assume that each active protein (P_i^*) can affect (depending on its functional class) the activity state of the other proteins with which it interacts (see Fig. 1); active activators and receptors activate their interaction partners, and active deactivators deactivate their interaction partners. Effectors are not allowed to act on any of the other proteins that are part of the network. As such, the activators and deactivators in the model loosely correspond to kinases and phosphates, whereas effectors would correspond to proteins that mediate a physiological function (e.g. transcription factors or proteins binding a transporter protein to facilitate its opening). To capture such physiological effects, we include a final protein in the network, an 'output' protein P_{out}, which is either activated or inhibited by the effector. We monitor the concentration of this protein in the presence of a ligand (i.e. signal) to quantify network dynamics. The ligand is assumed to act only on the receptor, either activating or deactivating it. Note that the receptors are modelled as activators, following a large body of observation that most natural receptors are kinases themselves or first interact with a kinase (modelling receptors as deactivators produce results similar to those shown in Figs 2 and 3, data not shown).

The interactions among the proteins result from a randomly generated network topology and allow us to write ordinary differential equations that describe the concentration of each of the proteins in the network. We assume bimolecular reactions resulting in equations of the form:

$$\frac{d[P_i^*]}{dt} = [P_i] \cdot \left(k_{ji} \cdot [P_j^*] + \alpha_i \cdot [L] + a_i\right) - [P_i^*] \cdot \left(k_{mi} \cdot [P_m^*] + d_i\right) \quad (1)$$

Equation (1) gives the rate of change in the active concentration of protein i (which is assumed to be a receptor for illustrative purposes) that is activated

© 2010 THE AUTHORS. J. EVOL. BIOL. doi:10.1111/j.1420-9101.2010.02101.x
JOURNAL COMPILATION © 2010 EUROPEAN SOCIETY FOR EVOLUTIONARY BIOLOGY

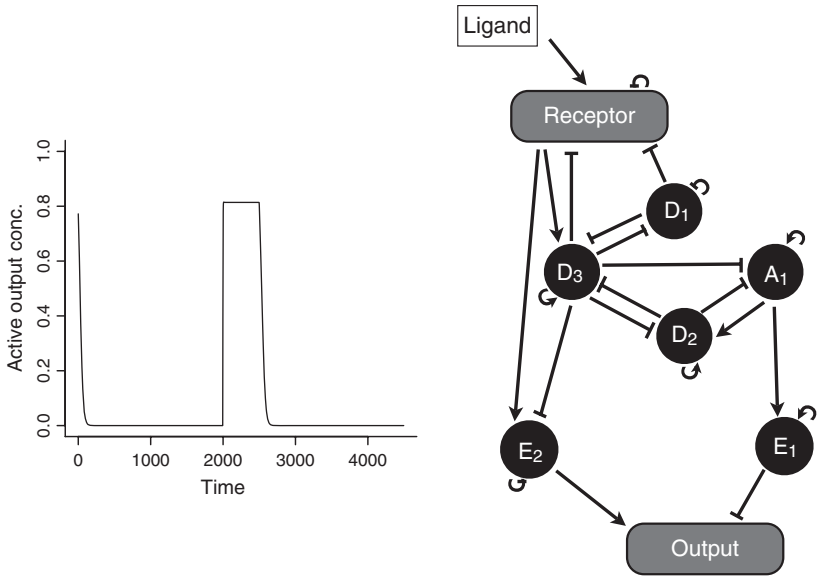

Fig. 1 Cartoon representation of a sample network (right) and its response to an incoming signal (left). The network model consists of proteins from four functional classes. Receptors relay the signal to intermediary proteins that they activate. These intermediary proteins can be activators (A) or deactivators (D) of other proteins. Effectors interact only with an output protein, whose active form is considered to mediate a physiological response. Each signalling protein is assumed to have an intrinsic self-activation (or deactivation). See Methods for further model details. Note that calculating network response involves monitoring the concentration of active form of each protein in the presence of a signal. The signal is introduced well after the system reaches initial steady state (at time 2000) and is removed after 500 time steps.

(deactivated) by protein j (m) and the signal (i.e. ligand). The interaction coefficients k_{ji} and k_{mi} denote the kinetic rates for the reactions mediated by the respective proteins, $[L]$ is the ligand concentration and α_i is the kinetic rate for ligand-based activation (or deactivation) of protein i. Note that for proteins other than the receptor, there will be no ligand effect. The coefficients a_i and d_i denote the rates for the unimolecular relaxation processes involving protein i. We assume that proteins relax to either their active or their inactive state, but not to both (i.e. $a_i \cdot d_i = 0$). We assume the total concentration of each protein to be constant and set it to one for computational ease (i.e. $[P_i] = 1 - [P_i^*]$).

In summary, the presented network model captures the basic biochemistry of signalling networks and allows us to derive time-response to a signal for a given model. A model consists of the numbers of proteins coming from different functional classes, the parameters controlling kinetic and relaxation activity of each protein and a network topology, defining the set of interactions between these proteins. For each generated model, the rate coefficient α_i is drawn randomly from a uniform distribution in the interval [−1.0, 1.0], the coefficients a_i and d_i are drawn randomly from a uniform distribution in the interval [−0.1, 0.1] and the interaction coefficients are drawn randomly from a uniform distribution in the interval [0, 1]. The low rate for self-reactions reflects the general observation that these reactions occur much more slowly compared to reactions mediated by other proteins (see for example (Porter & Armitage, 2002)).

Models similar to the one presented here have been used to analyse the dynamics of signalling networks (Binder & Heinrich, 2002; Heinrich *et al.*, 2002; Eungdamrong & Iyengar, 2004; Soyer *et al.*, 2006) and simulate their evolution (Azevedo *et al.*, 2006; Soyer & Bonhoeffer, 2006; François & Siggia, 2008; Troein *et al.*, 2009). More particularly, modelling biological systems with the above-given bimolecular reaction as the basic element is common, with several examples available in the modelling literature of signalling (e.g. Binder & Heinrich, 2002; Heinrich *et al.*, 2002; Kholodenko, 2006; Behar *et al.*, 2007a,b) and genetic(e.g. Wagner, 2000; Siegal & Bergman, 2002) networks. However, it must be noted that this reaction scheme ignores complex formation and multi-site phosphorylation. Despite this, the network model used here can display all of the dynamics that has been observed in real signalling networks (Soyer *et al.*, 2006).

Network dynamics and fitness

Duplication of a single gene (or its loss when considering whole-genome duplications), whose protein product is involved in a signalling network, will alter the dynamics of such a network and consequently the response of the cell to a given signal. Such alteration of signalling dynamics can have consequences at organism level, altering fitness (and phenotype) (Kollmann *et al.*, 2005; Peisajovich *et al.*, 2010). To link changes in dynamics to organismal fitness, we first need to quantify the former.

© 2010 THE AUTHORS. *J. EVOL. BIOL.* doi:10.1111/j.1420-9101.2010.02101.x
JOURNAL COMPILATION © 2010 EUROPEAN SOCIETY FOR EVOLUTIONARY BIOLOGY

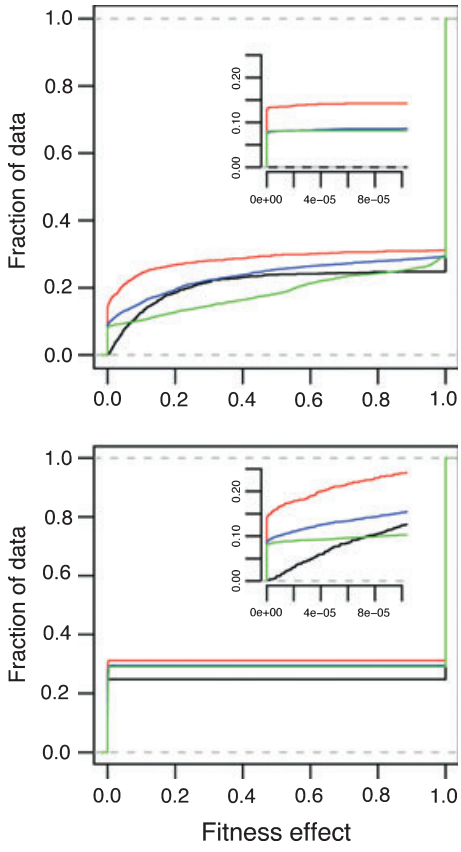

Fig. 2 Fitness effect (s) of gene duplications in each of the functional classes, receptor (black), activator (red), deactivator (blue) and effector (green). Data are shown as empirical cumulative distributions; each vertical line represents fraction of duplications that had a fitness effect shown on the x-axis. Results shown in top and bottom panels are obtained by assuming strong (sigma = 0.1) and weak selection (sigma = 100) on network dynamics, respectively (see eqn 4). The inset on each panel shows the distribution for the evolutionarily more relevant fitness ranges. Note that duplications that caused network instability are assigned the maximum fitness effect (of one). Data are compiled from 1000 random networks with connectivity, c = 0.5.

It has been argued (Heinrich *et al.*, 2002), and more recently shown experimentally(Sasagawa *et al.*, 2005; Peisajovich *et al.*, 2010), that some of the most relevant features of signalling dynamics relate to (i) the steady state activity of the network prior to a signal, (ii) the response amount and duration in the presence of a signal and (iii) the steady state activity post-signal. Here, we derive a measure for network dynamics based on these features as described below. In previous work, we and others have used similar measures to analyse the evolution of signalling networks under parasite interference (Salathé & Soyer, 2008) and to understand the key parameters underlying specific dynamics (Heinrich *et al.*, 2002).

To obtain response dynamics (D) for a given network, we first set $[P_i] = [P_i^*] = [P_i^{\text{tot}}]/2$ for all proteins in the network and $[L]$ = 0. We equilibrate the system by integrating the set of differential equations resulting from (eqn 1) for 2000 iterations. At the end of this period, we check whether the system has reached steady state using an eigenvalue analysis. If stability is reached, we record the active output protein concentration as the presignal steady state of the system, $[P^*_{\text{out}}]^{\text{SS}}_{\text{pre}}$. We then introduce a signal by setting $[L]$ = 1 and integrate the system for 500 iterations, after which the signal disappears (i.e. $[L]$ = 0). We then continue the integration for another 2000 iterations and again check for system stability. If the system is stable, we record the active output protein concentration as the post-signal steady state of the system, $[P^*_{\text{out}}]^{\text{SS}}_{\text{post}}$. Finally, we measure the response of the network by recording the change in the concentration of the active output protein during the time interval starting with the introduction of the signal and until the time point where the system first reaches post-signal steady state ($t^{\text{SS}}_{\text{post}}$), normalized by the maximum possible response. The exact calculation of the dynamic response of a network to an incoming signal, r, is given by

$$r = \frac{\sum_{t=2000}^{t^{\text{SS}}_{\text{post}}} \left| [P^*_{\text{out}}]^{\text{SS}}_{\text{pre}} - [P^*_{\text{out}}]_t \right|}{t^{\text{SS}}_{\text{post}} - 2000} \qquad (2)$$

With these measurements, we can write network dynamics as

$$D = r + [P^*_{\text{out}}]^{\text{SS}}_{\text{pre}} + [P^*_{\text{out}}]^{\text{SS}}_{\text{post}} \qquad (3)$$

As mentioned above, any change in D upon the duplication of a signalling protein might alter the fitness of the organism. To measure such fitness effects, s, we use

$$s = 1 - e^{-\frac{d(D,D')}{\sigma}} \qquad (4)$$

where D' stands for the dynamics obtained after duplication (or loss of one gene copy after WG duplication). Function d returns the sum of the absolute difference between current and original steady state values and the response. For duplications (or loss of a gene after whole-genome duplication) that lead to the network becoming unstable, we assume a fitness effect of one (i.e. we set $e^{-\frac{d(D,D')}{\sigma}} = 0$ for such systems). The parameter σ in eqn (4) allows us to control the fitness effects of any shift in network dynamics. Lower values of σ would mean that the shape of the network response is closely coupled to fitness, and any shift in dynamics have a large fitness cost. A biological example for this case would be the signalling network controlling bacterial chemotaxis, where effector concentration must remain in a tight interval for proper chemotaxis(Cluzel *et al.*, 2000; Kollmann *et al.*, 2005). Conversely, when σ is large, even very big shifts in the network dynamics would not alter fitness. A biological example would be a switch-like response in a transcription regulator, where only the

© 2010 THE AUTHORS. J. EVOL. BIOL. doi:10.1111/j.1420-9101.2010.02101.x
JOURNAL COMPILATION © 2010 EUROPEAN SOCIETY FOR EVOLUTIONARY BIOLOGY

presence of a response might matter and not its duration or post-signal level. Note that even under large σ, unstable networks are assigned a fitness of zero, because we assume that the dynamics of the network is still relevant for the organism. For example, the network should still be able to generate a response to an incoming signal, even though the timing and duration of such response might not matter. By performing the analysis under different values of σ, we explore how gene duplications are tolerated in these two scenarios (see Fig. 3).

Analysis of duplication effects

To analyse the effect of single duplications, we first generate random network models. Although these networks cannot be expected to capture all the intricacies of real networks, they are shown to be capable of displaying most of the dynamics seen in real biological networks (Soyer *et al.*, 2006). As discussed below, we further check for the possibility of our results being biased owing to the use of random network structures by analysing networks evolved *in silico*. To generate a random network, we first pick a random number of activators, deactivators and effectors, limiting the maximum number of proteins from any functional class to six for computational reasons. We then connect these with a receptor, an output protein and among themselves in a random fashion, obeying the limitations given above (as discussed in the main text, using multiple receptors in the generated networks do not alter the general conclusions made here). During this step, we use a given probability, c, for generating each connection, resulting in a corresponding average connectivity in the resulting networks. Second, we set the parameters governing the rate of kinetic (k_{ij}'s and α_i if i is a receptor) and relaxation (a_i or d_i) processes for a given protein by drawing random numbers from a uniform distribution in the interval [0, 1] and [0, 0.1], respectively. This modelling choice represents the general belief that self-relaxation process of proteins occurs much slower than their activation or deactivation mediated by other proteins. If the so-resulting network model is viable (i.e. has stable dynamics as explained earlier) and produces a response to an incoming signal above a given threshold (set to 0.1), we accept the network, otherwise we restart the process.

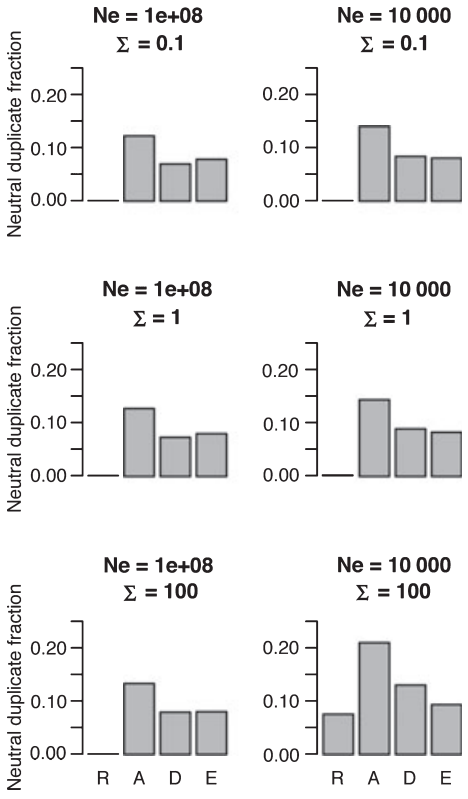

Fig. 3 The fraction of gene duplications that are neutral in each of the functional classes, receptor (R), activator (A), deactivator (D) and effector (E). Different panels correspond to different assumptions regarding the effective population size (N_e) and sigma values (as shown in panel headings). Data are compiled from the same 1000 random networks shown in Fig. 2.

We generate 1000 viable random networks for c = 0.3, 0.5, 0.7 and 1. For each network, we analyse the effect of duplicating each one of its proteins one by one, except the output protein. To model duplications, we simply add a new protein to the system, which is an exact copy of the one that is being duplicated. Note that this is equivalent of doubling the concentration of the protein involved. To model the loss of one copy of a gene after whole-genome duplication, we halve the total concentration of the involved protein. Such modelling of duplications explicitly assumes haploidy. In diploidy (or larger ploidy), we would expect to have more severe effects of duplication (in any functional class) on dynamics as this would correspond to a larger perturbation in the parameters of the model. As such, the observed increase in the number of neutral duplicates with decreasing population size might be an overestimate if this decrease in population size is associated with increasing ploidy.

To quantify the fraction of duplicates that could be considered neutral, we count the number of duplications with fitness effect below $1/2N_e$, where N_e gives the effective population size. We use different values for sigma and $2N_e$, with the latter based on realistic estimates (Lynch & Conery, 2003).

Evolved vs. random networks

Whereas the high numbers of sampled random networks should give an unbiased view of duplicates' effects on network dynamics, it is possible that evolved networks behave significantly differently from random networks.

© 2010 THE AUTHORS. *J. EVOL. BIOL.* doi:10.1111/j.1420-9101.2010.02101.x
JOURNAL COMPILATION © 2010 EUROPEAN SOCIETY FOR EVOLUTIONARY BIOLOGY

To check for this possibility, we generated 500 random networks and analysed duplicates' effects as before. We then simulated evolution of these networks and re-analysed duplicates' effects (as averaged over the entire population) at the end of these simulations. The Supporting Information Fig. S3 summarizes the results of this analysis for networks with different connectivity. Although these analyses are not conclusive, they suggest that our findings from random networks are extendable to evolved networks.

The evolutionary simulation of networks followed earlier approaches (Azevedo *et al.*, 2006; Soyer & Bonhoeffer, 2006). In brief, for each of the 500 random networks, we generated a homogenous population consisting of 500 identical copies of that network (i.e. the original network acted as a founder for the population). We then simulated evolution of this population for 1000 generations. At the end of each generation, a new population is produced from the current one using random drawing with replacement. A random individual is picked from the population and is cloned into the new population with a probability proportional to its fitness. Then, it is put back into the current population and a new draw is made, and the process continued until the new population contains 500 individuals. During replication of individual networks, mutations can occur with a probability of 0.001 per network and result in a small change (sampled from a normal distribution with mean zero and standard deviation one) in the kinetic parameters of a randomly selected protein. During evolution, network fitness, w, was defined by the distance of its dynamics to that of the founder network as before (i.e. $w = e^{-\frac{d(D,D')}{\sigma}}$). In other words, networks were evolved under stabilizing selection for response dynamics of the founder network. The parameter σ controls the strength of selection and was set to one for these simulations.

Compilation and analysis of genomic data

Definitions of gene families from 371 genomes across all three domains of life were retrieved from the eggNOG database (version 1) (Jensen *et al.*, 2008) (see Supporting Information Data S1 for a complete list of genomes used). The eggNOG database contains precalculated gene families for various taxonomic levels, for instance 'metazoa' or 'vertebrates'. For the purpose of this analysis, the gene families that spanned all three domains of life were used, i.e. clusters of orthologous groups (COG) and nonsupervised orthologous groups (NOG). These two are non-overlapping data sets; the first are based on a seed set of manually annotated gene families, and the second are the rest of the genes classified automatically into different orthologous groups. See Jensen *et al.* (Jensen *et al.*, 2008) for more details.

The annotation of the gene families in eggNOG was searched with appropriate keywords (see Supporting Information Table S1) to identify genes of the categories 'receptors', 'effectors' and 'signallers'. The latter category is taken to correspond to the activators and deactivators of the model. The resulting gene families are then further examined, and any families containing genes that are not clearly involved in signalling are purged. This manual curation involved picking sample genes from each retrieved gene family and going through their functional annotation given in the InterPro database (Mulder *et al.*, 2008). The final resulting database contained 699 gene families for receptors, 293 gene families for signallers and 69 gene families for effectors (see Supporting Information Data S2–S4 for a complete list of each). Any species that is completely missing genes from one of the three functional classes is removed from further analysis, resulting in the final data set spanning 293 species (see Supporting Information Data S1).

The same approach was taken to analysing a second data set focussed on groups in the eukaryotes that have differing effective population sizes. For this analysis, all fungal (the fuNOGS) and vertebrate (the veNOGS) gene families were retrieved from the eggNOG database (version 2) (Jensen *et al.*, 2008). These were compiled from 14 and 28 genomes, respectively. To minimize the effect of poor annotation in some genomes, only genomes with > 80% of their genes assigned to a gene family in each phyla were retained. This resulted in the retention of 7 fungal and 27 vertebrate genomes (see Supporting Information Data S5). Numbers of receptor, signaller and effector genes were calculated using the same technique as before, and all gene families that were assigned to more than one category were discarded (see Supporting Information Data S5 for species used and the genome and gene class size for each).

We used these data sets to compile the number of genes in a given functional class in each genome. This number is then normalized by the number of genes in the corresponding genome. We used the Wilcoxon rank-sum test (with continuity correction) as implemented in the statistical package 'R' (http://www.r-project.org/) to assess the effect of smaller effective population sizes on the expansion of receptor gene families. The distribution of receptors from eukaryotes is compared to the values from prokaryotes and on a more fine-grained level between fungi and vertebrates. These comparisons represent extremes of the scale of effective population size (Lynch & Conery, 2003) across domains and within eukaryotes, respectively. The same analysis is repeated on this data set using an alternative normalization scheme and also on another data set, which contained more specific data for prokaryotes (see Results and discussion).

Results and discussion

To quantify dynamical effects of duplications in a systematic fashion, we use a realistic model of signalling networks (see Methods). In particular, we generate

signalling networks consisting of a receptor, an 'output' protein, and a set of activators, deactivators and effectors, each modelled as two-state proteins (i.e. active, inactive). The receptor is coupled to an external signal, which can enhance or inhibit its activity, and effectors act on the output protein. The cascade of reactions mediated by the activators and deactivators relay the signal from the receptor to the effectors, which can either activate or deactivate the output protein. This model allows us to monitor the temporal changes in the concentration of the active form of each protein in the network in the presence of a signal. Hence, we can derive the response dynamics for a given network model consisting of a connectivity structure (i.e. network topology) and kinetic parameters (see Fig. 1).

To analyse the effects of gene duplications on network dynamics, we first generate random networks. For each network, we first derive the 'wild-type dynamics' and then duplicate each protein in the network one by one, recording the disruption caused in network dynamics (see Methods). The fitness effects of such disruption will depend on the importance of maintaining a given network dynamics. Here, we capture this dependency using a particular function, whose shape is tunable by a single parameter, sigma (see eqn 4). A high sigma value would correspond to a situation where the network dynamics is not relevant for fitness, i.e. the organism is not under selection for the exact dynamics of the network. Conversely, a small sigma would indicate that network dynamics is closely coupled to fitness, and any shift in dynamics would have a high fitness cost (e.g. the chemotaxis system described in the Introduction). Figure 2 shows the cumulative distribution of duplicates' fitness effects obtained from 1000 random networks and calculated for two different sigma values. Independent of the sigma value used, we find that a large fraction (approximately 70%) of the duplications result in the networks becoming unstable (i.e. network dynamics do not reach steady state at the end of simulation time), shown as a fitness effect of one.

From an evolutionary point, the important part of the data presented in Fig. 2 is the lower end of the cumulative distributions, where the duplication resulted in an effect of nearly zero. Theory suggests that for a nonbeneficial mutation to possibly fix in the population, it has to have a fitness effect lower than a critical value in the order of $1/2N_e$, where N_e corresponds to effective population size (Gillespie, 2004). Although it is difficult to measure N_e, estimates suggest that it is generally $> 10^8$ for prokaryotes and in the range of 10^4–10^6 for invertebrates and vertebrates (Lynch & Conery, 2003). Given these estimates, we analyse the fraction of duplicates that resulted in fitness effects below $1/2N_e$. Figure 3 summarizes the results for different values of N_e and sigma. Interestingly, we find that a higher fraction of duplications are neutral for intermediary proteins, in particular activators, in comparison with receptors. For receptors, neutral duplications are almost nonexistent. This result makes intuitive sense; duplication of a receptor would have a direct and strong effect on network dynamics as all incoming signals have to pass through the receptors, while effects of duplicating intermediary proteins could be dampened by the overall network structure (and dynamics). In other words, receptor duplications would bear a higher fitness cost because of error propagation through the network. In line with this view, we find that using random networks, where each network contains multiple receptors that can sense and relay signals in different ways, results in an increase in the fraction of neutral duplicates for receptors (see Supporting Information Fig. S1). Biologically, such a 'multiple receptors' model would correspond to cross-talk among different networks (i.e. signals). Although cross-talk seems to be exploited by the cell in certain cases (McClean *et al.*, 2007), most signals are believed to be processed by isolated networks, and several mechanisms for avoiding cross-talk are documented (Alves & Savageau, 2003; Behar *et al.*, 2007a,b; Csikász-Nagy *et al.*, 2010). As such, we concentrate here on the 'one receptor per network' model.

We find that the fraction of neutral receptor duplications becomes detectable only when we assume a low N_e ($=10^4$) combined with a high sigma ($=100$) (see Fig. 3). This scenario corresponds to signalling networks in organisms with small population size (e.g. vertebrates) and for which the exactness of dynamics is not important for organism fitness. Although these conditions result in neutral fraction of duplicates to increase for any functional class, we find this effect to be most pronounced for receptors. This is because of the differences in the distribution of duplicates' fitness effects for the different functional classes (see inset, Fig. 2). To check for the effect of network connectivity (the ratio between existing and all possible interactions in a network) on these results, we constructed random networks with different average connectivity, c. As expected, we find that increasing c results in a decrease in neutral duplicates. The more reactions an average protein participates in, the less likely it is for its duplicate to have a small effect on the network dynamics. In the extreme case of fully connected networks, there are no neutral duplication events any more (see Supporting Information Fig. S2). Interestingly, the fact that most neutral duplicates result from the intermediary proteins remains unaffected by network connectivity, although it is most pronounced for low to medium connectivity. This main result seems to strengthen when we consider networks evolved under stabilizing selection (see Methods) rather than randomly generated networks. As shown in Supporting Information Fig. S3, analysis of such evolved networks gives qualitatively the same results as the analysis of random networks, indicating that the effects of duplication events are not biased by the model structure.

© 2010 THE AUTHORS. *J. EVOL. BIOL.* doi:10.1111/j.1420-9101.2010.02101.x
JOURNAL COMPILATION © 2010 EUROPEAN SOCIETY FOR EVOLUTIONARY BIOLOGY

The single duplication events we considered so far are only one way of generating duplicate gene copies. Another major event is the duplication of entire genomes (Aury *et al.*, 2006; Semon & Wolfe, 2007), resulting in double copies of each gene in the organism. This is an intrinsically neutral event in terms of network dynamics as doubling the concentration of each protein in a signalling network would not alter its dynamics. However, any subsequent loss-of-function mutations would possibly result in disruption of the network dynamics. Here, we model such events (i.e. gene copy loss after a WGD) by halving the concentration of each of the genes in a given network. As before, we generate 1000 random networks and repeat the procedure for each gene in each network. Surprisingly, we find results highly similar to single duplication case. As shown in Supporting Information Fig. S4, the fraction of neutral gene copy losses is highest for activators, followed by deactivators and effectors. Again, events involving receptors are rarely tolerated.

To summarize, this theoretical analysis shows that response dynamics would constrain neutral fixation of duplication (and gene copy loss after a WGD) events in receptors more strongly than in intermediary proteins of a signalling network. More importantly, we find that the distribution of fitness effects of receptor duplications has a significantly different shape (see Fig. 2) than that found for other signalling proteins. As a result, we find neutral fixation of duplications in receptors is possible only in organisms with small population size and in signalling networks where exactness of dynamical response is not crucial. It is possible to extrapolate from this prediction that an expansion of receptor numbers could occur only in organisms with small effective population size or that have undergone multiple rounds of WGD events. This is difficult to test as the actual protein family sizes in different organisms would be determined by several factors including rate of duplication and nature of selective forces acting on duplicates (and on the organism). Further, the theoretical analysis presented here is only relevant for cases where early fixation of duplicates is driven through genetic drift and does not account for potential adaptive fixation events. Determining which

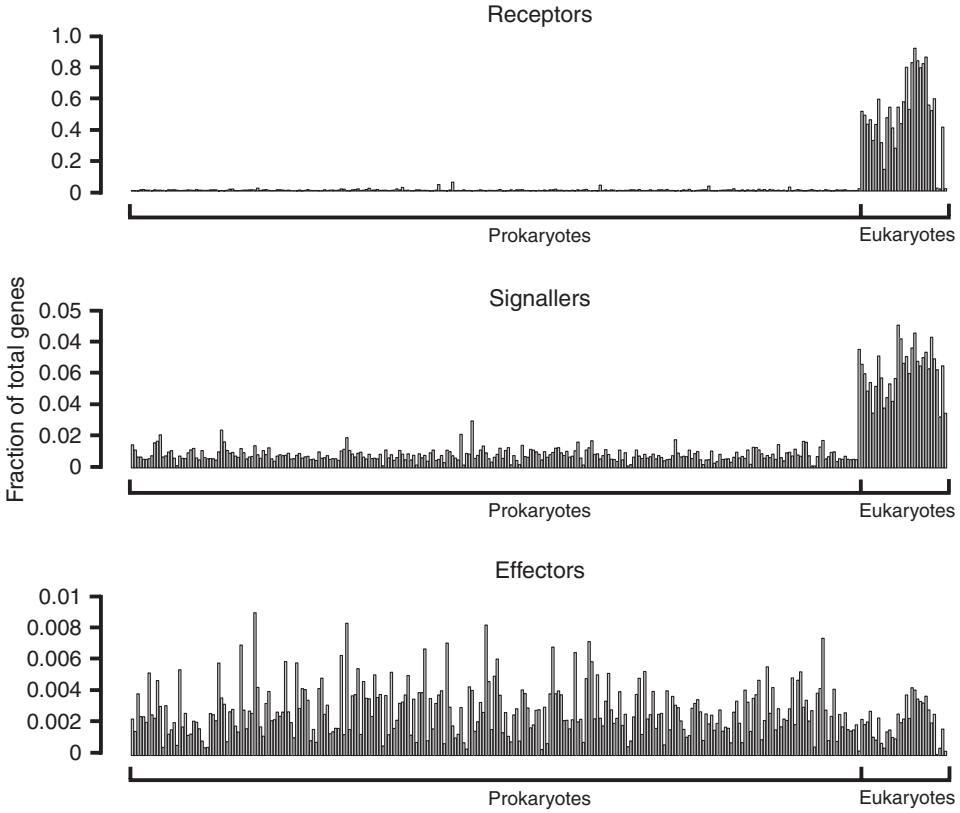

Fig. 4 The distribution of the fraction of genes in the receptor, effector and signaller gene families, over all species analysed. Panels from top to bottom show the fraction of genes in a given genome that is coding for functional families receptor, signaller and effector. On the *x*-axis, we have all analysed species, with eukaryotes ordered to the right. The distribution of these values for the eukaryotes was compared to the distribution for the prokaryotes. This analysis shows that eukaryotic genomes harbour a significantly higher fraction of receptors compared to prokaryotes (Wilcoxon rank-sum test: $W = 8317$, $P < 2.2e{-}16$). The same observation is also significant in the case of signallers (Wilcoxon rank-sum test: $W = 8352$, $P < 2.2e{-}16$) but not for the effectors that showed the opposite trend (Wilcoxon rank-sum test: $W = 3007$, $P = 0.009$).

© 2010 THE AUTHORS. *J. EVOL. BIOL.* doi:10.1111/j.1420-9101.2010.02101.x
JOURNAL COMPILATION © 2010 EUROPEAN SOCIETY FOR EVOLUTIONARY BIOLOGY

mode of fixation applies to different proteins is very difficult, if not impossible, further confounding any empirical analysis. Despite these difficulties, we analysed the family size of different protein families involved in signalling to see whether there would be any indication of expansion of number of receptors in organisms with small population size.

We compiled a data set of all signalling protein families from 371 species with fully sequenced genomes. The final data set contained 65 592 proteins spanning 1061 gene families in 293 species (see Methods for data compilation and analysis). We classified these proteins based on their annotated function as receptors, effectors and signallers, where the latter class corresponds to deactivators and activators of the model. Using eukaryotes and prokaryotes as two ends of the scale of predicted effective population size (Lynch & Conery, 2003), we find that species with smaller effective population sizes (the eukaryotic genomes) harbour a significantly larger fraction of receptors compared to species with larger effective population sizes (the prokaryotic genomes) as shown in Fig. 4. The same observation holds for signallers but not for effectors (but see also Supporting Information Figs S5 and S6).

There are several possible caveats with this empirical analysis. It is possible for example that annotation of the genomes is incomplete or biased. Even for the fully annotated genomes, the annotations can be erroneous. Further, both our classification of signalling proteins and the use of keywords to retrieve genes belonging to such functional classes may be incomplete and crude. We have tried several approaches to reduce the possible effects of such caveats. First, we have used an alternative normalization scheme with the above data set and normalized the data by the total number of genes involved in signalling in a given genome (rather than by the total number of genes in that genome). As both the number of genes in each family and the total number of signalling genes result from the same analysis, such normalization might give a more reliable comparison among different genomes, reducing any effects from biased or incomplete annotations. Using such normalization, we still find eukaryotes to harbour significantly more receptors than prokaryotes (Supporting Information Fig. S5). Secondly, we have used an alternative data set, which specialized on signalling proteins in bacteria (Galperin, 2005). This manually curated data set contained all signalling proteins in bacterial genomes and presents possibly the best resource for such proteins over all sequenced bacterial genomes. In particular, this data set lists the following functional gene families in each of the analysed genomes: histidine kinases, methyl-accepting receptors, adenylate cyclases, response regulators, Tyr-specific protein kinases, proteins with phosphodiesterase activity and proteins involved in the turnover of secondary messengers. Following the described activities of these proteins (Galperin, 2005), we classified the first three classes of genes as receptors, the response regulators as effectors and the remaining genes as signallers. To further refine this classification, we used the information in the same data set in the presence of transmembrane (TM) regions in these proteins. In particular, we classified histidine kinases with TM regions as receptors, and those without as signallers. We then combined this bacterial data set with the data we compiled on eukaryotic genomes and analysed the resulting data set as before. This analysis shows that eukaryotic genomes harbour significantly more receptors compared to prokaryotes (Supporting Information Fig. S6).

Finally, we carried out a more fine-grained analysis between the unicellular eukaryotes (represented by fungi) and multicellular eukaryotes (represented by vertebrates) to see whether the same trend held within domain as across (see Supporting Information Fig. S7). These two groups represent two extremes of effective population size in eukaryotes. The results mirrored those of the across-domain analysis (i.e. eukaryotes vs. prokaryotes), showing there had been a significantly larger expansion (Wilcoxon rank-sum test: $W = 0$, $P = 6.238e{-}05$) of receptor genes in the vertebrates when compared to the fungi, whereas there was no significant difference in the proportion of signaller genes (Wilcoxon rank-sum test: $W = 91$, $P = 0.89$) or effector genes (Wilcoxon rank-sum test: $W = 103.5$, $P = 0.71$).

Conclusions

Here, we analysed the initial fate of a duplicate in the context of a signalling network. In particular, we quantified the effect on response dynamics when genes from different functional classes in a signalling network are duplicated (or lost after a WGD). We find that most duplications in all functional classes cause strong disruptions in network dynamics. Considering fitness effects of such disruptions in network dynamics, we find that only a small fraction of gene duplications can fix through genetic drift (i.e. neutrally). Among all functional classes considered, receptors have the lowest chance for neutral fixation (and neutral gene copy loss after a WGD). As expected, the fraction of duplications that can fix neutrally increases in all functional classes with decreasing population size and selective pressures on network dynamics. Interestingly, this effect is most pronounced for receptors. Such a differential effect of decreased population size on the neutral fixation of duplicates might manifest itself as an expansion of receptors in species with small population size (i.e. vertebrates) or in those that have undergone multiple WGDs.

In line with such a possibility, we find that eukaryotic genomes harbour more receptors compared to prokaryotes. Further, this possibility fits well with more specific analyses of signalling proteins; it has been observed that

© 2010 THE AUTHORS. *J. EVOL. BIOL.* doi:10.1111/j.1420-9101.2010.02101.x
JOURNAL COMPILATION © 2010 EUROPEAN SOCIETY FOR EVOLUTIONARY BIOLOGY

G-protein-coupled receptors are selectively maintained following WGDs (Semyonov *et al.*, 2008), and protein kinases, which would loosely correspond to activators in the presented model, are overrepresented in mouse (Forrest *et al.*, 2003). It is important to note, however, that these empirical analyses cannot be taken as proof of the model findings. This is because the empirically observed patterns (e.g. expansion of receptors in eukaryotes) can have a variety of causes, including both adaptive and neutral processes. Although disentangling these causes requires a much more in-depth analysis, the presented model indicates that neutral processes can have a significant contribution.

This work concentrates on the initial fixation of a duplicate through genetic drift at network level. As such, its findings do not exclude possible cases of positive and negative selection in the retention (i.e. fixation and subsequent divergence) of duplicates in signalling networks, which can arise from redundancy under certain conditions (Nowak *et al.*, 1997; Wagner, 2000; Salathé & Soyer, 2008) or from dosage effects (Cook *et al.*, 1998; Papp *et al.*, 2003; Aury *et al.*, 2006; Hakes *et al.*, 2007). The presented analysis provides a null hypothesis for the expected number of signalling proteins in different organisms based on neutral fixation alone. As such, it is conceptually similar to previous analyses concentrating on the effects of neutral processes on genome (Lynch & Conery, 2003) and network complexity (Soyer & Bonhoeffer, 2006). In particular, the former analysis indicates that larger genome size observed in eukaryotes is a result of decrease in population size, leading to higher instance of duplicate retention. This view is extended to signalling networks in this work, resulting in the finding that decreasing population size can affect duplicates from different functional classes in these networks differently.

As noted before (Lynch, 2007), models focusing on neutral processes provide the right context to evaluate findings from high-throughput and system-level studies. Furthermore, the extension of the presented model and the data analysis can be used to detect selective deviations from neutral expectations as has been carried out at sequence level (Mustonen & Lässig, 2007).

Acknowledgments

We are thankful to David Liberles and Lars Juhl Jensen for fruitful discussions. O. S. S. acknowledges the support of Exeter University, Science Strategy. C. J. C. acknowledges support from the Science Foundation Ireland (SFI) Stokes Lectureship Programme (Reference number: 07/SK/B1236A).

References

Alves, R. & Savageau, M.A. 2003. Comparative analysis of prototype two-component systems with either bifunctional or monofunctional sensors: differences in molecular structure and physiological function. *Mol. Microbiol.* **48**: 25–51.

Artzy-Randrup, Y., Fleishman, S.J., Ben-Tal, N. & Stone, L. 2004. Comment on ''Network motifs: simple building blocks of complex networks'' and ''Superfamilies of evolved and designed networks''. *Science* **305**: 1107. author reply 1107.

Aury, J.M., Jaillon, O., Duret, L., Noel, B., Jubin, C., Porcel, B.M., Ségurens, B., Daubin, V., Anthouard, V., Aiach, N., Arnaiz, O., Billaut, A., Beisson, J., Blanc, I., Bouhouche, K., Càmara, F., Duharcourt, S., Guigo, R., Gogendeau, D., Katinka, M., Keller, A.M., Kissmehl, R., Klotz, C., Koll, F., Le Mouel, A., Lepère, G., Malinsky, S., Nowacki, M., Nowak, J.K., Plattner, H., Poulain, J., Ruiz, F., Serrano, V., Zagulski, M., Dessen, P., Bètermier, M., Weissenbach, J., Scarpelli, C., Schachter, V., Sperling, L., Meyer, E., Cohen, J. & Wincker, P. 2006. Global trends of whole-genome duplications revealed by the ciliate Paramecium tetraurelia. *Nature* **444**: 171–178.

Azevedo, R.B., Lohaus, R., Srinivasan, S., Dang, K.K. & Burch, C.L. 2006. Sexual reproduction selects for robustness and negative epistasis in artificial gene networks. *Nature* **440**: 87–90.

Behar, M., Dohlman, H.G. & Elston, T.C. 2007a. Kinetic insulation as an effective mechanism for achieving pathway specificity in intracellular signaling networks. *Proc. Natl. Acad. Sci. USA* **104**: 16146–16151.

Behar, M., Hao, N., Dohlman, H.G. & Elson, T.C. 2007b. Mathematical and Computational Analysis of Adaptation via Feedback Inhibition in Signal Transduction Pathways. *Biophys. J.* **93**: 806–821.

Binder, H. & Heinrich, R. 2002. Dynamic stability of signal transduction networks depending on downstream and upstream specificity of protein kinases. *Mol. Biol. Rep.* **29**: 51–55.

Brenner, S.E., Hubbard, T., Murzin, A. & Chothia, C. 1995. Gene duplications in H. influenza. *Nature* **378**: 140.

Cluzel, P., Surette, M. & Leibler, S. 2000. An ultrasensitive bacterial motor revealed by monitoring signaling proteins in single cells. *Science* **287**: 1652–1655.

Cook, D.L., Gerber, A.N. & Tapscott, S.J. 1998. Modeling stochastic gene expression: implications for haploinsufficiency. *Proc. Natl. Acad. Sci. USA* **95**: 15641–15646.

Csikász-Nagy, A., Cardelli, L. & Soyer, O.S. 2010. Response dynamics of phosphorelays suggest their potential utility in cell signaling. *J. R. Soc. Interface*, doi: 10:1098/rsif.2010.0336.

Deutschbauer, A.M., Jaramillo, D.F., Proctor, M., Kumm, J., Hillenmeyer, M.E., Davis, R.W., Nislow, C. & Giaever, G. 2005. Mechanisms of haploinsufficiency revealed by genome-wide profiling in yeast. *Genetics* **169**: 1915–1925.

Duarte, J.M., Cui, L., Wall, P.K., Zhang, Q., Zhang, X., Leebens-Mack, J., Ma, H., Altman, N. & dePamphilis, C.W. 2006. Expression pattern shifts following duplication indicative of subfunctionalization and neofunctionalization in regulatory genes of Arabidopsis. *Mol. Biol. Evol.* **23**: 469–478.

Eungdamrong, N.I. & Iyengar, R. 2004. Modeling cell signaling networks. *Biol. Cell* **96**: 355–362.

Evangelisti, A.M. & Wagner, A. 2004. Molecular evolution in the yeast transcriptional regulation network. *J. Exp. Zool. B Mol. Dev. Evol.* **302**: 392–411.

Force, A., Lynch, M., Pickett, F.B., Amores, A., Yan, Y.L. & Postlethwait, J. 1999. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151**: 1531–1545.

Forrest, A.R., Ravasi, T., Taylor, D., Huber, T., Hume, D.A. & Grimmond, S. 2003. Phosphoregulators: protein kinases

and protein phosphatases of mouse. *Genome Res.* **13**: 1443–1454.
François, P. & Siggia, E.D. 2008. A case study of evolutionary computation of biochemical adaptation. *Phys. Biol.* **5**: 26009.
Galperin, M.Y. 2005. A census of membrane-bound and intracellular signal transduction proteins in bacteria: bacterial IQ, extroverts and introverts. *BMC Microbiol.* **5**: 35.
Gillespie, J.H. 2004. *Population Genetics A Concise Guide*. The Johns Hopkins University Publisher, Baltimore.
Gu, X., Zhang, Z. & Huang, W. 2005. Rapid evolution of expression and regulatory divergences after yeast gene duplication. *Proc. Natl. Acad. Sci. USA* **102**: 707–712.
Hakes, L., Pinney, J.W., Lovell, S.C., Oliver, S.G. & Robertson, D.L. 2007. All duplicates are not equal: the difference between small-scale and genome duplication. *Genome Biol.* **8**: R209.
He, X. & Zhang, J. 2005. Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. *Genetics* **169**: 1157–1164.
Heinrich, R., Nell, B.G. & Rapoport, T.A. 2002. Mathematical models of protein kinase signal transduction. *Mol. Cell* **9**: 957–970.
Hughes, T. & Liberles, D.A. 2007. The pattern of evolution of smaller-scale gene duplicates in mammalian genomes is more consistent with neo- than subfunctionalisation. *J. Mol. Evol.* **65**: 574–588.
Huminiecki, L. & Wolfe, K.H. 2004. Divergence of spatial gene expression profiles following species-specific gene duplications in human and mouse. *Genome Res.* **14**: 1870–1879.
Jensen, J.L., Julien, P., Kuhn, M., von Mering, C., Muller, J., Doerks, T. & Bork, P. 2008. eggNOG: automated construction and annotation of orthologous groups of genes. *Nucleic Acids Res.* **36**: D250–D254.
Kholodenko, B.N. 2006. Cell-signalling dynamics in time and space. *Nat. Rev. Mol. Cell Biol.* **7**: 165–176.
Kollmann, M., Løvdok, L., Bartholomé, K., Timmer, J. & Sourjik, V. 2005. Design principles of a bacterial signalling network. *Nature* **438**: 504–507.
Landry, C.R., Castillo-Davis, C.I., Ogura, A., Liu, J.S. & Hartl, D.L. 2007. Systems-level analysis and evolution of the phototransduction network in Drosophila. *Proc. Natl. Acad. Sci. USA* **104**: 3283–3288.
Lynch, M. 2007. The evolution of genetic networks by non-adaptive processes. *Nat. Rev. Genet.* **8**: 803–813.
Lynch, M. & Conery, J.S. 2000. The evolutionary fate and consequences of duplicate genes. *Science* **290**: 1151–1155.
Lynch, M. & Conery, J.S. 2003. The origins of genome complexity. *Science* **302**: 1401–1404.
Lynch, M., Sung, W., Morris, K., Coffey, N., Landry, C.R., Dopman, E.B., Dickinson, W.J., Okamoto, K., Kulkarni, S., Hartl, D.L. & Thomas, W.K. 2008. A genome-wide view of the spectrum of spontaneous mutations in yeast. *Proc. Natl. Acad. Sci. USA* **105**: 9272–9277.
McClean, M.N., Mody, A., Broach, J.R. & Ramanathan, S. 2007. Cross-talk and decision making in MAP kinase pathways. *Nat. Genet.* **39**: 409–414.
Moore, R.C. & Purugganan, M.D. 2003. The early stages of duplicate gene evolution. *Proc. Natl. Acad. Sci. USA* **100**: 15682–15687.
Mulder, N.J., Kersey, P., Pruess, M. & Apweiler, R. 2008. *In silico* characterization of proteins: UniProt, InterPro and Integr8. *Mol. Biotechnol.* **38**: 165–177.
Mustonen, V. & Lässig, M. 2007. Adaptations to fluctuating selection in Drosophila. *Proc. Natl. Acad. Sci. USA* **104**: 2277–2282.
Nowak, M.A., Boerlijst, M.C., Cooke, J. & Smith, J.M. 1997. Evolution of genetic redundancy. *Nature* **388**: 167–171.
Papp, B., Pál, C. & Hurst, L.D. 2003. Dosage sensitivity and the evolution of gene families in yeast. *Nature* **424**: 194–197.
Peisajovich, S.G., Garbarino, J.E., Wei, P. & Lim, W.A. 2010. Rapid diversification of cell signaling phenotypes by modular domain recombination. *Science* **328**: 368–372.
Porter, S.L. & Armitage, J.P. 2002. Phosphotransfer in Rhodobacter sphaeroides chemotaxis. *J. Mol. Biol.* **324**: 35–45.
Salathé, M. & Soyer, O.S. 2008. Parasites lead to evolution of robustness against gene loss in host signaling networks. *Mol. Syst. Biol.* **4**: 202.
Sasagawa, S., Ozaki, Y., Fujita, K. & Kuroda, S. 2005. Prediction and validation of the distinct dynamics of transient and sustained ERK activation. *Nat. Cell Biol.* **7**: 365–373.
Semon, M. & Wolfe, K.H. 2007. Consequences of genome duplication. *Curr. Opin. Genet. Dev.* **17**: 505–512.
Semyonov, J., Park, J.I., Chang, C.L. & Hsu, S.Y. 2008. GPCR genes are preferentially retained after whole genome duplication. *PLoS ONE* **3**: e1903.
Siegal, M.L. & Bergman, A. 2002. Waddington's canalization revisited: developmental stability and evolution. *Proc. Natl Acad. Sci. USA* **99**: 10528–10532.
Sopko, R., Huang, D., Preston, N., Chua, G., Papp, B., Kafadar, K., Snyder, M., Oliver, S.G., Cyert, M., Hughes, T.R., Boone, C. & Andrews, B. 2006. Mapping pathways and phenotypes by systematic gene overexpression. *Mol. Cell* **21**: 319–330.
Soyer, O.S. & Bonhoeffer, S. 2006. Evolution of complexity in signaling pathways. *Proc. Natl. Acad. Sci. USA* **103**: 16337–16342.
Soyer, O.S., Salathé, M. & Bonhoeffer, S. 2006. Signal transduction networks: topology, response and biochemical processes. *J. Theor. Biol.* **238**: 416–425.
Tirosh, I. & Barkai, N. 2007. Comparative analysis indicates regulatory neofunctionalization of yeast duplicates. *Genome Biol.* **8**: R50.
Troein, C., Locke, J.C., Turner, M.S. & Millar, A.J. 2009. Weather and seasons together demand complex biological clocks. *Curr. Biol.* **19**: 1961–1964.
Vogel, C. & Chothia, C. 2006. Protein family expansions and biological complexity. *PLoS Comput. Biol.* **2**: e48.
Wagner, A. 1994. Evolution of gene networks by gene duplications: a mathematical model and its implications on genome organization. *Proc. Natl. Acad. Sci. USA* **91**: 4387–4391.
Wagner, A. 2000. The role of population size, pleiotropy and fitness effects of mutations in the evolution of overlapping gene functions. *Genetics* **154**: 1389–1401.
Wagner, A. 2005. Energy constraints on the evolution of gene expression. *Mol. Biol. Evol.* **22**: 1365–1374.
Walsh, J.B. 1995. How often do duplicated genes evolve new functions? *Genetics* **139**: 421–428.
Zhang, J. 2003. Evolution by gene duplication. *Trends Ecol. Evol.* **18**: 292–298.

Supporting information

Additional Supporting Information may be found in the online version of this article:

© 2010 THE AUTHORS. *J. EVOL. BIOL.* doi:10.1111/j.1420-9101.2010.02101.x
JOURNAL COMPILATION © 2010 EUROPEAN SOCIETY FOR EVOLUTIONARY BIOLOGY

Appendix S1 Legends to Figures S1 to S7.
Figure S1 The fraction of gene duplications that are neutral in each of the functional classes, receptor (R), activator (A), deactivator (D) and effector (E) when using a model that allows for multiple receptors per network (see main text).
Figure S2 The fraction of gene duplications that are neutral in each of the functional classes, receptor (R), activator (A), deactivator (D) and effector (E).
Figure S3 The fraction of gene duplications that are neutral in each of the functional classes, before (dark grey bars), i.e. for randomly generated networks, and after *in silico* evolution (light grey bars).
Figure S4 The fraction of gene loss events that are neutral in each of the functional classes, receptor (R), activator (A), deactivator (D) and effector (E).
Figure S5 The distribution of the fraction of genes in the receptor, effector and signaller gene families, over all species analysed and using total number of genes involved in signalling as a normalization factor.
Figure S6 The distribution of the fraction of genes in the receptor, effector and signaller gene families, over all species analysed (using a dedicated data set for bacterial genomes as explained in the main text).
Figure S7 Proportion of receptor, signaller and effector gene families in fungal and vertebrate genomes.
Table S1 Keywords used to identify gene families from each of the three categories used.
Data S1 List of genomes used in the empirical analysis presented in Figure 4.
Data S2 List of genes classified as 'effector' for the empirical analysis.
Data S3 List of genes classified as 'receptor' for the empirical analysis.
Data S4 List of genes classified as 'signaller' for the empirical analysis.
Data S5 List of genomes used in the empirical analysis presented in Figure S7.

As a service to our authors and readers, this journal provides supporting information supplied by the authors. Such materials are peer-reviewed and may be re-organized for online delivery, but are not copy-edited or typeset. Technical support issues arising from supporting information (other than missing files) should be addressed to the authors.

Received 25 May 2010; revised 30 July 2010; accepted 2 August 2010

© 2010 THE AUTHORS. *J. EVOL. BIOL.* doi:10.1111/j.1420-9101.2010.02101.x
JOURNAL COMPILATION © 2010 EUROPEAN SOCIETY FOR EVOLUTIONARY BIOLOGY